Cognition 146 (2016) 110–120

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Oops, scratch that! Monitoring one's own errors during mental calculation

Ana L. Fernandez Cruz [a,b,*], Santiago Arango-Muñoz [c,d], Kirsten G. Volz [a]

[a] Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Tübingen 72076, Germany
[b] McGill University Integrated Program in Neuroscience, Montréal H3A 2B4, Canada
[c] Institute for Philosophy II, Ruhr University, Bochum 44780, Germany
[d] Grupo Conocimiento, Filosofía, Ciencia, Historia y Sociedad, Instituto de Filosofía, Universidad de Antioquia, Colombia

ARTICLE INFO

Article history:
Received 10 July 2014
Revised 23 July 2015
Accepted 6 September 2015
Available online 24 September 2015

Keywords:
Error monitoring
Error detection
Feeling of error
Confidence
Metacognition
Metacognitive feeling
Number bisection task

ABSTRACT

The feeling of error (FOE) is the subjective experience that something went wrong during a reasoning or calculation task. The main goal of the present study was to assess the accuracy of the FOE in the context of mental mathematical calculation. We used the number bisection task (NBT) to evoke this metacognitive feeling and assessed it by asking participants if they felt they have committed an error after solving the task. In the NBT participants have to determine whether the number presented in the middle of a triplet corresponds to the arithmetic mean of the two outer numbers (e.g., 07_16_25) with a Yes/No answer. Our results show that FOE reports were strongly correlated with arithmetic errors and numerical properties of the NBT, suggesting that the FOE accurately represents the error. This finding indicates that even very fast metacognitive feelings are reliable when it comes to evaluating one's own mental performance. Moreover, our results suggest that the occurrence of FOEs is determined by the fluency with which each triplet was solved and the post-decision evaluation processes that occurred after the NBT was solved. Additionally, we asked participants to report their confidence in the given answer for the cases where they did not report FOEs. Participants reported less confidence for the (objectively) incorrect answers than for the (objectively) correct ones, suggesting that in cases where they did not have a conscious FOE they still were able to implicitly detect their errors. Remarkably, confidence was also determined by the fluency of the NBT.

1. Introduction

When solving math problems such as multiplication or division, people sometimes get the (gut) feeling that their calculations have gone wrong and that, therefore, they should not endorse the output of their mental calculation. This feeling appears as a spontaneous phenomenal experience that points to the fact that the calculation might be mistaken and motivates the reasoner to revise what she has been doing. Everyday observation suggests that this phenomenon is not restricted to the classroom; it generalizes to all contexts where people carry out mental actions, such as making mental rotations when calculating their way from one point to another by using a map, deciding between two possible actions, reasoning about the probability of an event, or mentally calculating how much money they spent in the last week. In situations like these, people sometimes report experiencing a "feeling of error" (henceforth FOE) that alerts them about a possible mistake in their mental processing. The subjective experience that something went wrong is assumed to arise during or right after the mental action and is fundamental for further correction and improvement in calculating and reasoning.

The FOE has been classified as a metacognitive or epistemic feeling in the literature on metacognition (Arango-Muñoz, 2014; Gangemi & Bourgeois-Gironde, 2014; Thompson & Johnson, 2014). Accordingly, the FOE is conceived as a phenomenal experience directed toward a mental state, process or disposition, that motivates certain behaviors such as changing the strategy or checking the outcome of a mental action (for an overview see Arango-Muñoz & Michaelian, 2014; De Sousa, 2009; Moulin & Souchay, 2013). Metacognitive feelings are particularly interesting because they make people aware of mental conditions that they would not notice in the absence of such feelings. For instance, in the case of the tip-of-the-tongue phenomenon, the feeling points to the agent that she is in possession of a piece of information although she has no access to it in her memory, and so motivates the individual to keep trying to remember (Brown & McNeill,

* Corresponding author at: Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Tübingen 72076, Germany.
E-mail address: ana.fernandezcruz@mail.mcgill.ca (A.L. Fernandez Cruz).

http://dx.doi.org/10.1016/j.cognition.2015.09.005
0010-0277/

1966; Schwartz, 2001; Schwartz & Metcalfe, 2011; see Brown, 2012 for a recent review).

Most of the empirical studies working on metacognitive feelings and reasoning have focused on positive feelings whereby the person detects a correct answer. For example, Boekaerts and Rozendaal (2010) assessed students' confidence in a mathematical task where students had to report their confidence on a 10 point scale before and after they had produced the solution to two types of mathematical problems: computation and application problems.[1] Predominantly, they found effects of the type of mathematical problem (computation and application problem) and the time of measurement (before or after solving the problem) on the accuracy of the confidence. Similarly, the feeling of rightness (henceforth FOR) has been addressed by Thompson, Prowse Turner, and Pennycook (2011); they investigated the FOR after individuals solved conditional reasoning and syllogistic problems. Participants had to provide an initial, intuitive response to the reasoning problem, as well as a retrospective evaluation of their intuitive answer based on their FOR. The authors reported a negative correlation between the FOR and the reaction time of the initial intuitive response, such that fluent processing (as indicated by shorter reaction times) was associated with a higher FOR.

On the other hand, the studies addressing performance monitoring by negative feelings, like the feeling of error, have followed two different traditions. One focuses on error detection and error awareness of bodily actions (see Wessel, 2012 for a review), and the other focuses on the metacognitive feeling of error related to reasoning (Gangemi & Bourgeois-Gironde, 2014; Thompson & Johnson, 2014). The first tradition uses behavioral paradigms such as the Go/No-Go (Dhar, Wiersema, & Pourtois, 2011; Murphy, Robertson, Allen, Hester, & O'Connell, 2012), the flanker task (Hughes & Yeung, 2011; Scheffers & Coles, 2000) and the antisaccade paradigm (Endrass, Franke, & Kathmann, 2005; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001) to study the detection and awareness of erroneous bodily movements. Often, these researchers are also interested in determining whether the electrophysiological indices of cortical error processing (i.e., error related negativity (ERN) or the error positivity (Pe)) are associated with error awareness (Boldt & Yeung, 2015; Steinhauser & Yeung, 2010). In contrast, the second tradition, focuses on the detection and awareness of erroneous mental reasoning episodes, and therefore uses logical, probabilistic and mathematical reasoning tasks (De Neys, 2012; De Neys, Cromheeke, & Osman, 2011; De Neys & Glumicic, 2008; Gangemi & Bourgeois-Gironde, 2014).

These two traditions have developed their own methods, paradigms and models in parallel without any interaction. One of the aims of this paper is to bridge this gap by integrating elements from the two perspectives. On the one hand, following the error detection tradition, we chose a speeded button press task (Murphy et al., 2012; Rabbitt, 1966) in which a mathematical task was embedded. On the other hand, the mathematical task was chosen to evoke reasoning errors, as has been done in the metacognitive error awareness literature (De Neys, 2012; Gangemi & Bourgeois-Gironde, 2014). These factors allowed a close examination of participants' behavior related to error monitoring, thereby merging the two traditions. The novelty of this approach is that it incorporates reasoning and/or mental calculation errors in the framework of error detection, which has traditionally focused on action or motor errors. Furthermore, following the metacognitive tradition, we additionally asked for introspective reports about the feelings that accompanied the task (Koriat, 2000, 2007; Reder & Ritter, 1992; Gangemi et al., 2014) and used established measures, such as the Gamma correlation, to assess the accuracy of the feelings (Koren, Seidman, Goldsmith, & Harvey, 2006; Nelson, 1984). In line with both traditions, we also asked participants to rate their confidence (Boldt & Yeung, 2015; Scheffers & Coles, 2000; Yeung & Summerfield, 2012).

With this integrative goal in mind, we considered three specific aims and three hypotheses. First, we wanted to determine the accuracy of the FOE. Based on previous studies on metamemory (Koriat, 2000; Paynter, Reder, & Kieffaber, 2009; Reder & Ritter, 1992) and metareasoning studies (Gangemi & Bourgeois-Gironde, 2014), we hypothesized that: (1) the FOE in a reasoning task is a reliable signal of error (as is the case in error detection on motor tasks). That is, we expect that participants would mainly report having a FOE after having committed a mistake in their calculation.

Second, we were interested in defining the determinants of the FOE. Two main factors that have been proposed in the literature as determinants of metacognitive feelings were considered with this goal: fluency and post-decision evaluation. In the tradition of metacognitive studies, fluency refers to the ease with which a piece of information is processed and/or comes to the mind (Oppenheimer, 2008; Schwarz, 2010), for example, the speed with which an item is retrieved from memory (Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005, see Koriat, 2007 for a review). Accordingly, if fluency plays a role in the FOE reports, then there should be a higher probability of no-FOE reports for fluent calculations, and higher probability of FOE reports for disfluent calculations (Jiang & Hong, 2014; Thompson, 2009; Thompson et al., 2011). Therefore, we hypothesized that: (2.1) if the FOE is determined by fluency processing, participants will report less FOEs on fluent trials as compared to disfluent trials. The second factor that has been proposed as a determinant of metacognitive feelings is post-decision evaluation process. Metacognitive tradition on error monitoring specifies that this process occurs after acting or making a decision and serves in evaluating the likelihood that the decision or action will result in a favorable or unfavorable outcome (Vickers & Lee, 1998, 2000). According to Vickers and Lee (1998, 2000), metacognitive feelings of confidence and error are the product of an accumulator system that progressively and continuously accumulates and evaluates evidence in favor of or against the initial response (see Yeung & Summerfield, 2012 for a review). In other words, participants keep considering the problem and checking their answer after giving the answer to the problem, and this post-evaluation leads to error detection and subsequent behavioral slowdown, even in the absence of feedback (Rabbitt, 1966). The behavioral slowdown is not restricted to the primary task (e.g., the mathematical task in our experimental design), but it can generalize and affect other immediately following tasks, as has been demonstrated by recent studies (Cho, Orr, Cohen, & Carter, 2009; Forster & Cho, 2014; Notebaert & Verguts, 2011). Thus, given that error detection is normally followed by a slowdown in the subsequent behavior, the reaction time of the FOE report (the task that immediately follows the mathematical calculation in our experimental design) can be used as an index of a continued post-decision evaluation process. Based on these considerations, we hypothesized that: (2.2) if the FOE is determined by a post-decision evaluation, we expect to find that participants take longer to report whether or not they had a FOE in the mathematical task after having committed an error, compared to when they had made no error; we also expect participants to take longer to report FOEs than to report no-FOEs.

Our third and last aim was to explore the extent to which participants were sensitive to their missed errors, that is, participants' sensitivity to the errors they fail to report (i.e., after no-FOE report).

[1] An example of an application problem: "The Mount Everest has the highest mountaintop on earth. Its height is 8848 m above sea level. The lowest point of the earth's crust is in the Pacific Ocean at 11,034 m below sea level. What is the difference between the highest and the lowest points on earth?". An example of a computation problem: "68.2 − 4.73 = . . .?".

We explored this possibility by considering that, although participants commit errors and often overlook them (i.e., do not provide a FOE report), they may have different degrees of awareness concerning their mistakes. For example, they might report low confidence after not reporting a FOE. There is a growing literature that points to the fact that subjects implicitly detect or are slightly aware of some of their "unreported" errors (De Neys, 2012, 2014; De Neys & Glumicic, 2008; De Neys et al., 2011). To test whether participants had some degree of awareness of their missed errors, we asked participants to report their confidence in their answer when they reported not having a FOE. Thus, we hypothesized that: (3) if participants had some degree of awareness of their missed errors, we expect that they will report lower confidence for missed errors (i.e., reporting a no-FOE after an erroneous response) than for appropriate rejections (i.e., reporting a no-FOE after a correct response).

To test our three hypotheses we adapted the number bisection task (henceforth NBT), a task that has been used in neuropsychological studies to assess quantitative capabilities in number processing. Particularly, we used the verification version[2] of the NBT so as to elicit and test the FOE (material was kindly provided by Nuerk and colleagues, cp. Nuerk et al., 2002): First, participants were briefly presented with number triplets (e.g., 07_16_25) and then had to quickly decide whether or not the number in the middle of a triplet was the arithmetic mean of the two outer numbers by a Yes/No answer (Moeller et al., 2011; Nuerk et al., 2002). We manipulated participants' response window so as to increase their likelihood of errors in the NBT and the uncertainty about the correctness of their answers. Specifically, the response window was individually adapted to participants' speed so that it amounted to the mean of the last 5 triplets (gliding window) and maximally amounted to 2.5 s. After each trial, participants had 2 s to report whether or not they had a FOE. This manipulation aimed at producing feeling-based assessments (Koriat, 2007; Reder & Ritter, 1992) of their performance in the NBT.

2. Methods

2.1. Participants

Thirty right-handed volunteers (15 female, mean age: 24 years; SD = 2.9, range: 19–30) participated in the experiment for a monetary compensation of 12 € per hour. Given that the main goal of our study was to induce FOEs during mathematical reasoning, we did not recruit participants who could be assumed to be mathematically inclined, that is, we excluded engineers, mathematicians or physicists, for whom the NBT might have been too easy. Data from two participants had to be excluded from the analyses since they did not follow the instructions, i.e., they skipped over all the questions about the FOE without answering.

2.2. Stimuli and design

We used 200 triplets from the NBT by Nuerk et al. (2002). These triplets were selected from a set of originally 360 triplets based on difficulty measures from previous studies. Namely, we chose the 25 most difficult triplets of each of the 8 categories (see below), based on the reaction times of participants for solving each triplet from a study by Moeller, Wood, Doppelmayr, and Nuerk (2010). This selection was done to increase the likelihood of evoking FOEs in the present study.

Triplets were categorized into bisectable and non-bisectable triplets following Nuerk et al. (2002). On the one hand, bisectable triplets correspond to triplets in which the middle number is the arithmetic mean of the two outer numbers, e.g., 12_21_30. Non-bisectable triplets, on the other hand, are triplets in which the middle number is not the arithmetic mean, e.g., 19_25_29. Participants' performance on the NBT was evaluated using four mathematical factors – two for each triplet's category – that have been shown to influence participants' reaction times (fluency) and accuracy while solving the NBT (Moeller et al., 2011; Nuerk et al., 2002). We thus propose that if fluency is a decisive factor for the FOE reports, fluent triplets – as determined by the factors listed below – would elicit less FOE reports compared to disfluent triplets, which would elicit more FOE reports. The two factors considered for the bisectable triplets were *multiplicativity* and *range*. The factor *multiplicativity* specified whether the triplet belonged to a multiplication table, e.g., 16_20_24, or not, e.g., 22_26_30. Triplets that belong to multiplication table are henceforth labeled as multiplicative triplets, and triplets that do not belong as non-multiplicative ones. Previous studies (Moeller et al., 2010; Nuerk et al., 2002) reported non-multiplicative triplets as being more difficult, i.e., have greater error percentages and longer reaction times than multiplicative ones. We thus expected these triplets to be less fluent than multiplicative triplets. The factor *range* indicates if the numerical distance between the outer numbers of the triplet is small, ranging from 4 to 8, e.g., 16_18_20 (distance between the outer numbers is 4), or large, ranging between 12 and 18, e.g., 14_21_28 (distance between the outer numbers is 14). Based on results from Moeller et al. (2010) and Nuerk et al. (2002) that reported non-multiplicative triplets with a large range to have greater error rates and longer reaction times, we expected these triplets to be less fluent than multiplicative ones with shorter ranges.

On the other hand, the factors considered for the non-bisectable triplets were *bisection possibility* and *distance to mean*. The *bisection possibility* accounted for the possibility that the outer numbers of the triplet could *in principle* have an integer as arithmetic mean, that is, the outer numbers have an integer as mean but this number is not displayed in the middle of the triplet. For instance, the triplet 41_56_57 (which has possibility of bisection) has an arithmetic mean of 49, whereas the (non-bisectable) triplet 12_20_21 has a decimal as arithmetic mean, namely 16.5. Triplets that cannot be bisected have been shown to be more fluent than triplets that could *in principle* have integers as means (Moeller et al., 2010; Nuerk et al., 2002). Finally, the *distance to mean* refers to the numerical distance between the number in the middle of the triplet and the actual correct arithmetic mean of the outer numbers. The distance between these numbers can be small or large. Small *distances to mean* are between 0.5 and 1.5, e.g., 25_31_35, which has a distance to the actual mean (30) of 1. Large distances are between 2 and 8, e.g., 47_58_59, which has distance of 5. In a nutshell, non-bisectable triplets that can *in principle* be bisectable and have smaller distances to means are expected to be less fluent than non-bisectable ones without bisection possibility and larger ranges, because the former display longer reaction times and greater error percentages than the latter (Moeller et al., 2010; Nuerk et al., 2002). Thus, the design totaled 4 categories (2 by 2 factorial design) for the bisectable and 4 categories for the non-bisectable triplets, hence 8 categories of triplets in total.

2.3. Procedure

After a detailed instruction about the NBT and the procedure of the experiment, participants worked on 10 practice trials to get

[2] The verification version of the NBT is answered by a yes–no answer, as opposed to the production version of the NBT in which participants only get the 2 outer numbers and have to produce and write down the correct arithmetic mean, if it exists for the corresponding trial (Moeller, Klein, Fischer, Nuerk, & Willmes, 2011; Nuerk, Geppert, van Herten, & Willmes, 2002).

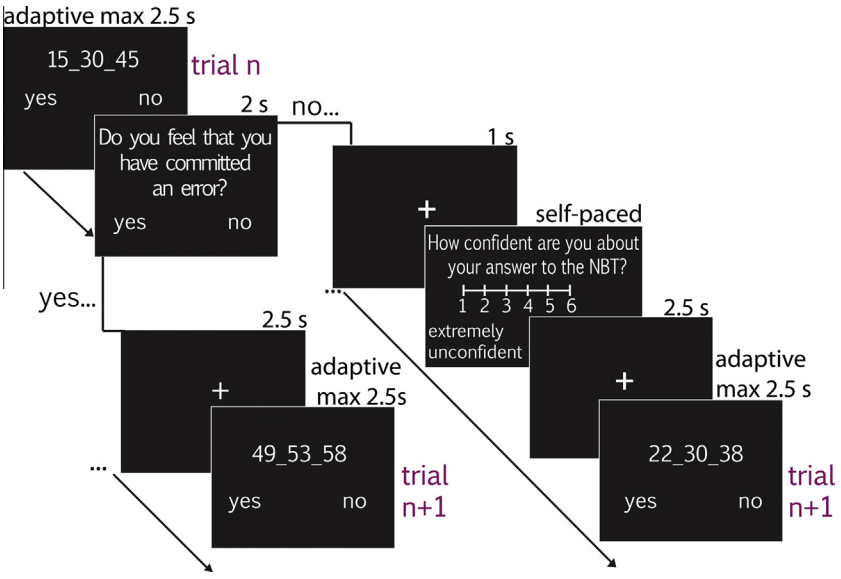

Fig. 1. Diagram of the experimental procedure. Participants were first presented with three numbers and they had to decide if the middle number of this triplet was also the integer arithmetic mean of the two outer numbers. They provided an answer by pressing Yes or No response keys. If they were too slow providing an answer a new trial started, i.e., another triplet was presented. In case participants provided an answer (yes/no), they were then asked if they felt they had committed an error on their calculation (FOE question). The trial progressed depending on participants' response: (a) if participants answered *they felt they had not committed an error*, they were asked how confident they were about their answer to the triplet using a 6 point Likert scale (see definition of the scale in Section 2.3), or (b) if participants answered that *they felt they had committed an error*, a new trial started and another triplet was presented.

used to the procedure and timing. The first 5 practice trials had no time limit and participants received feedback about the correctness of their answer after each trial. In contrast, the last 5 practice trials had the same time restrictions as in the actual experiment and no feedback was provided. Participants had the opportunity to ask comprehension questions after the practice trials before starting with the actual experiment. They were unaware of the purpose of the experiment and were only told that they were participating in a study investigating numerical cognition. During each trial, participants were presented with triplets of numbers and their task was to rapidly decide whether the number in the middle of the triplet was the arithmetical mean of the two outer numbers or not. (The experimental procedure is summarized in Fig. 1.) They pressed Yes or No answer keys (index and middle finger) assigned to the right hand to give their answers. The position (left/right) of the Yes/No answer keys and screen display was counterbalanced across participants. This was done maintaining position congruency (left/right) between the screen display and answer keys. The 200 triplets were individually randomized and were presented only once during the experiment. No performance feedback was given during the experiment. The time participants were given to answer each NBT triplet was adaptive and 2.5 s at maximum, i.e., it decreased over the course of the experiment if participants answered correctly and in less than 2.5 s. The actual time participants had for answering was calculated online trial-by-trial as the average of the time it took them to answer the last 5 triplets. If one or more of the last 5 triplets was answered incorrectly (or missed) a buffer time of 2.5 s was included to calculate the average of the current time, and if one or more of those 5 triplets was answered correctly then its actual response time was included for the average. The time of the first 5 triplets was calculated as the average of the response time of the 5 practice trials (within the time limit). This manipulation was done to increase the likelihood of errors and FOEs. Immediately after participants answered each of the triplets their error monitoring was tested by asking them to report whether they felt they had committed an error on their calculation or not (FOE question). Literally, the question asked was: "Do you feel that you have committed an error?" (Haben Sie das Gefühl, eine fehlerhafte Antwort gegeben zu haben?). Participants were instructed to answer this question as quickly as possible within a time limit of 2 s by pressing Yes/No answer keys assigned to the right hand (index and middle finger). This time limit was used to capture the spontaneity of the FOE, that is, we wanted participants to answer as quickly and spontaneously as possible instead of basing their reports on reflective thinking about their performance or further recalculations. If participants answered "NO" to the FOE question, they were asked to report their degree of confidence about their answer to the presented triplet in the NBT. Their confidence was assessed using a 6 point Likert-type scale without any time restrictions: 1 was defined as *extremely unconfident (aeusserst unsicher)* and 6 as *extremely confident (aeusserst sicher)*. The values 2–5 were defined as: 2 = *very unconfident (sehr unsicher)*, 3 = unc*onfident (unsicher),* 4 = *confident (sicher),* 5 = *very confident (sehr sicher)*. As soon as participants answered the confidence question, a new triplet was presented. If participants missed a triplet, i.e., took longer than 2.5 s to answer, the experiment continued automatically with the presentation of a new triplet and neither FOE nor confidence questions were asked.

3. Results

Dependent variables were tested for normal distribution using Kolmogorov–Smirnov (K–S) tests before conducting any analysis. All variables were significantly normal as demonstrated by the K–S test statistics (D): number of reported FOEs ($D(28) = 0.13$, $p > 0.05$), number of correctly ($D(28) = 0.11$, $p > 0.05$) and incorrectly answered triplets ($D(28) = 0.14$, $p > 0.05$), reaction time of the NBT ($D(28) = 0.08$, $p > 0.05$) and reaction time of the FOE report ($D(28) = 0.13$, $p > 0.05$), gamma ($D(28) = 0.12$, $p > 0.05$), confidence ($D(28) = 0.13$, $p > 0.05$) and percentage of error ($D(28) = 0.15$, $p > 0.05$).

3.1. Performance on the number bisection task

Given that participants had a maximum of 2.5 s to provide an answer to each triplet, 23.7% of the trials (i.e., 45 trials, SD = 24)

were missed on average per participant. The number of missed triplets was not significantly different for the bisectable (mean = 22.04 ± SD = 12) and non-bisectable triplets (mean = 23.07 ± SD = 13.4), $t(27) = 0.66$, $p = 0.51$). Naturally, the missed problems were not considered for the analysis since once a trial was missed neither FOE reports nor confidence assessments were collected. Participants answered three-quarters of the trials correctly (74.2% (SD = 9.8%)), which is significantly higher than the 50% chance level ($t(27) = 13.03$, $p < 0.001$, one-tailed). Note that misses were already excluded from the analyses, i.e., 100% corresponds exclusively to the number of answered trials. As revealed by a paired sample *t*-test, the reaction times of bisectable and non-bisectable triplets were not significantly different, ($t(27) = 0.59$, $p = 0.55$, two tailed). However, the number of bisectable triplets answered correctly (mean = 47.5 ± SD = 16) was significantly less than the number of non-bisectable triplets answered correctly (mean = 60.43, ± SD = 16.1), ($t(27) = 3.2$, $p < 0.001$, two tailed).

Correctly and incorrectly answered trials did not differ significantly as to response time, ($t(27) = -0.85$, $p = 0.39$, two-tailed). The mean reaction time (RT) ± standard deviation (SD) for all correct trials was: 1629.2 ± 194 ms and for all incorrect trials was: 1641.7 ± 234 ms. The reaction times did not differ either when both types of triplets, bisectable and non-bisectable, were considered separately. Mean RT ± SD for correct bisectable trials was: 1669.8 ± 221 ms and for incorrect bisectable trials was: 1639.2 ± 240 ms, ($t(27) = 1$, $p = 0.32$, two-tailed). Mean RT ± SD for correct non-bisectable trials was: 1626.9 ± 209 ms and for incorrect non-bisectable triplets was: 1661.1 ± 247 ms, ($t(27) = 1.218$, $p = 0.23$, two-tailed, see Fig. 2).

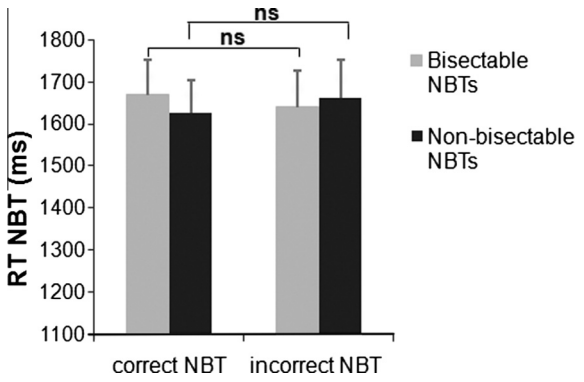

Fig. 2. Reaction time of the objectively correct and incorrect triplets. The reaction of the bisectable and non-bisectable triplets did not differ for the objectively correct or incorrect answered triplets. Ns = not statistically significant.

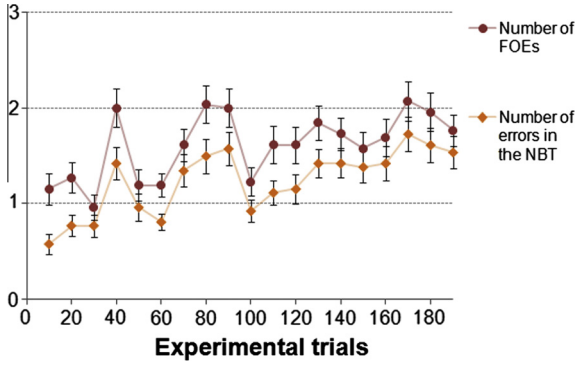

Fig. 3. The trajectory of the number of errors and reported FOEs over the experiment are plotted with every data point representing the average of 10 subsequent trials. Each point represents the mean value of 10 trials computed across subjects and errors bars represent the 95% confidence interval. Errors and feelings of error (FOEs) occurred all throughout the experiment without accumulating at the beginning of the experiment, demonstrating that the instructions were clear and participants understood the task.

3.2. FOE accuracy

Our first research question was to assess the accuracy of the FOE. Before doing so, we verified that our paradigm was successful in evoking reasoning errors and FOEs. Fig. 3 shows that the number of errors and reported FOEs occurred all through the 190 trials of the experiment, suggesting that our experimental manipulation was successful in eliciting reasoning errors that were not the product of comprehension errors at the beginning of the experiment. Participants reported, on average, a FOE in 21 ± 9% of the answered triplets (i.e., mean ± SD = 30 ± 15 FOEs on average per participant). Bisectable and non-bisectable triplets significantly differed in the number of FOEs reported; participants reported significantly more FOEs for the bisectable triplets, mean ± SD = 17.6 ± 9.9, than for the non-bisectable triplets, mean ± SD = 12.8 ± 7.3, ($t(27) = 2.7$, $p = 0.01$, two tailed).

We tested our first hypothesis, that the FOE is a reliable signal of error, by means of three different analyses: (1) Following the tradition in metacognitive studies (Koren et al., 2006; Nelson, 1984), we used Kruskal–Goodman Gamma correlation (γ) to measure metacognitive resolution. The metacognitive resolution indicates the degree to which FOE reports corresponded to errors and no-FOEs corresponded to correctly answered tasks, i.e., how these variables vary concurrently. Gamma measurements can reach a value between −1 and 1, where 0 represents no relationship between the variables or chance level, large negative values represent an inverse association, and high positive values a strong association between the variables. Gamma was calculated within participants using the following formula: γ = (Concordances − Discordances)/(Concordances + Discordances), wherein concordances were defined as the FOE hits, namely, cases when participants committed an error on the NBT and reported a FOE, and also the cases where no error occurred and no FOE was reported (correct FOE rejection). The discordances, on the other hand, were the cases in which no error was committed but a FOE was reported (FOE false alarm), and those in which an error was committed but there was not a FOE report (FOE omission). Fig. 4 shows the values (mean percentage of each trial category over the total number of experimental trials) that were used to calculate gamma.

As expected, the gamma correlation was very high, γ mean ± SD = 0.74 ± 0.17, and significantly larger than 0.5, as demonstrated by a one sample *t*-test, ($t(27) = 4.6$, $p < 0.001$, one sample). These results suggest a strong association between the calculation errors and the FOEs, as well as between the correct solved tasks and the no-FOEs reports. Gamma values were significantly different for bisectable vs. non-bisectable triplets; the gamma of the bisectable triplets was lower, γ mean ± SD = 0.67 ± 0.23, than the one calculated for the non-bisectable triplets, γ mean ± SD = 0.79 ± 0.17, ($t(27) = 3.5$, $p < 0.01$, two tailed).

Additionally, a Pearson's correlation analysis revealed that the number of reported FOEs significantly correlated with the number of incorrectly answered NBT ($r = 0.89$, $p < 0.001$, $n = 28$, see Fig. 5).

3.3. Determinants of the FOE

3.3.1. Fluency

3.3.1.1. Factors influencing fluency. Before testing if fluency was a determinant of the FOEs, we evaluated the participants' performance on the NBT to determine whether the mathematical factors reported in Nuerk and colleagues study (2002) (see Section 2.2) influenced the fluency of the response to the NBT in our experi-

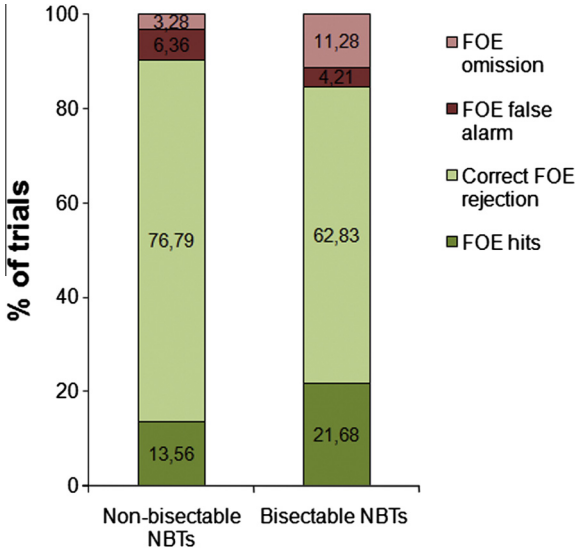

Fig. 4. Accuracy of the feeling of error (FOE). Mean percentages of the categories of trials that were defined according to the FOE reports and performance on the NBT. Categories of trials were defined as: (1) FOE hits: reported FOEs after incorrect responses to the NBT (dark green). (2) Correct FOE rejections: no FOEs were reported after correct responses (light green). (3) FOE false alarms: FOEs were reported after a correct response (dark red). (4) FOE omissions: FOEs were not reported after incorrect responses (light red). Note that these values were used to calculate the Kruskal–Goodman Gamma correlation.

mental paradigm. The fluency of the response was assessed according to participants' reaction time.

With this goal, we performed 2 by 2 within-subjects ANOVAs independently for bisectable and non-bisectable triplets using reaction time as dependent variable. For the bisectable triplets, the factors *range* (wide vs. small) and *multiplicativity* (multiplicative vs. non-multiplicative) were considered. *Bisection possibility* (possible vs. impossible) and *distance to mean* (large vs. small) were the factors used for the ANOVA of the non-bisectable triplets. The detailed results of this analysis are shown in the supplementary material given that it is a replication of previous results concerning the mathematical properties of the NBT (Nuerk et al., 2002), (see Section 5. Supplementary material: Mathematical factors and performance on the Number Bisection Task).

3.3.1.1.1. Bisectable triplets. Triplets with wider range were answered significantly slower than triplets with smaller range ($F(1,27) = 23.17$, $p < 0.001$) indicating that the factor range determined the fluency of the bisectable triplets. On the other hand, the factor *multiplicativity* did not show an effect on the fluency of the triplets.

3.3.1.1.2. Non-bisectable triplets. The factors *bisection possibility* ($F(1,27) = 36.42$, $p < 0.001$, partial $\eta^2 = 0.57$) and *distance to mean* ($F(1,27) = 4.60$, $p < 0.05$, partial $\eta^2 = 0.15$) had significant effects on the reaction time of the non-bisectable triplets. Participants took longer to answer the bisectable possible triplets and the triplets with smaller *distance to mean* (disfluent triplets) versus the triplets without possibility of bisection and those with larger *distance to mean*, respectively.

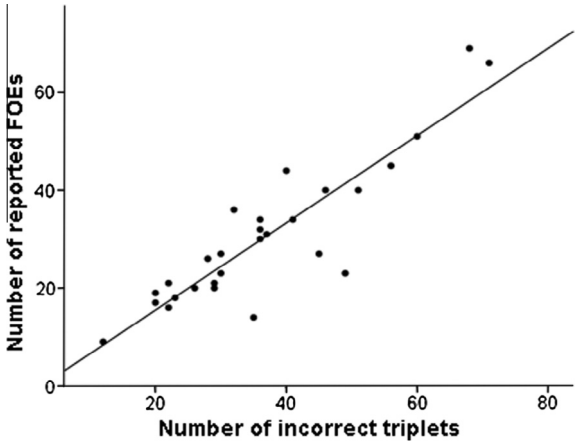

Fig. 5. Pearson correlation of the number of incorrect answered triplets and reported feelings of error ($r = 0.896$, $p < 0.001$, $n = 28$). Each point represents the averaged number of FOEs and incorrect triplets, per participant, over the 190 trials.

3.3.1.2. Fluency as a determinant of the FOE. After verifying that the fluency of the answers given to the NBT was determined by some of the factors reported by Nuerk and colleagues, we tested if disfluent trials, as determined by the significant factors detailed above, evoked more FOEs than fluent trials. We specifically performed separate ANOVAs with the factor *range* for the number of FOE reports that followed the bisectable triplets, and *distance to mean* and *bisection possibility* for the ones that followed non-bisectable triplets. Based on the literature on metacognition and fluency (Benjamin et al., 1998; Koriat & Ma'ayan, 2005; see Koriat, 2007 for a review), we expected that disfluent bisectable triplets, i.e., those with larger ranges would elicit more FOEs than fluent ones. Note that the fluency, i.e., reaction time, of the non-multiplicative and multiplicative triplets didn't differ significantly in our study. Similar results were expected for disfluent non-bisectable triplets, i.e., we expected that bisectable possible triplets and those with shorter *distance to mean* would elicit more FOEs than fluent non-bisectable ones.

We performed this analysis pooling together objectively correct and incorrect answered triplets. A one factor (wide vs. small) within subjects ANOVA revealed no main effects of the *range* ($F(1,27) = 1.01$, $p = 0.32$, partial $\eta^2 = 0.03$) on the number of the reported FOEs concerning the bisectable triplets. On the other hand, for the non-bisectable triplets, the number of reported FOEs was significantly larger for disfluent triplets with small *distance to mean*, mean ± SD = 4.2 ± 2.9, than for the fluent ones with larger *distance to mean*, mean ± SD = 2.2 ± 1.9, ($F(1,27) = 21.18$, $p < 0.001$, partial $\eta^2 = .44$). Similarly, *bisection possibility* had an effect on the number of reported FOEs. Disfluent triplets with possibility of bisection had a larger number of reported FOEs, mean ± SD = 3.6 ± 2.4, than the fluent ones with impossibility of bisection, mean ± SD = 2.8 ± 2.4, ($F(1,27) = 5.29$, $p < 0.05$, partial $\eta^2 = 0.16$, see Fig. 6 for a summary of the fluency effects on the number of reported FOEs). The interaction between the two factors due to (*distance to mean* * *bisection possibility*) was not significant for the number of FOEs ($F(1,27) = 0.87$, $p = 0.35$, partial $\eta^2 = 0.031$). Finally, in order to rule out that the fluency effects were not exclusively due to differences in error rates across conditions (i.e., due to a larger number of errors in the disfluent triplets vs. fluent ones), we performed the same analysis with the significant factors (*distance to mean* and *bisection possibility*) for correct and incorrect NBTs separately. The factor *distance to mean* showed a significant effect on the number of reported FOEs for correct ($F(1,27) = 16.13$, $p < 0.001$, partial $\eta^2 = 0.37$) and incorrect triplets ($F(1,27) = 13.53$, $p = 0.001$, partial $\eta^2 = 0.34$), indicating that fluency was indeed a determinant of the FOEs evoked in these triplets, irrespective of their correctness. On the other hand, the factor *bisection possibility* boarded on significance for correct triplets ($p = 0.092$). Its effect on the incorrect triplets was not significant.

3.3.2. Post-decision evaluation processes

To find out whether the reported FOEs were the product of continued post-decision evaluation processes, we calculated if the reaction time of the FOE reports were generally longer after

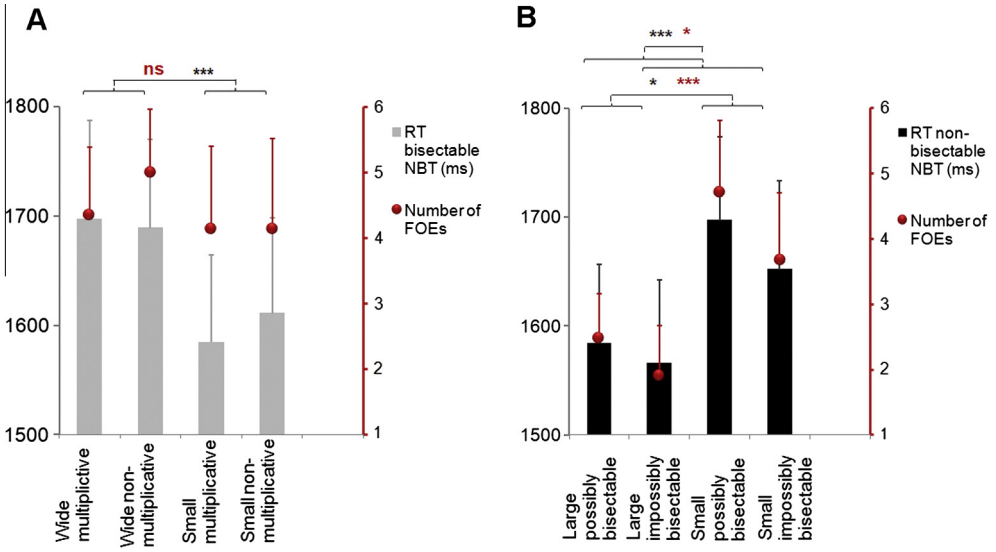

Fig. 6. Effects of the mathematical factors on the reaction time of the NBT and the number of FOEs. The graphs show fluency effects and their relationship with the number of FOEs for the bisectable triplets **(A)** and non-bisectable triplets **(B)** when objectively correct and incorrect triplets were averaged together. The left *y*-axes on each graph (A, B) show the range of values for the reaction time of the NBT (gray column bars). The right *y*-axes on each graph (A, B) show the range of values for the number of FOEs (red circles). Black stars indicate the significance level for differences in the reaction time of the NBT and red stars for the number of FOEs: $^{*}p < 0.05$; $^{***}p < 0.001$. Error bars indicate the upper half of the 95% confidence interval. Ns = not statistically significant.

an objective error was committed in the NBT compared to the FOE reports that followed a correct answer. In other words, we tested if post-error slowing evoked by incorrectly answered triplets was extended to the subsequent task, i.e., the FOE report. We then measured the reaction time of the FOE reports in dependence of the occurrence of a FOE. Specifically, we calculated one way within subjects ANOVAs on the reaction time of the FOE reports for the bisectable and non-bisectable triplets. The factor considered for these ANOVAs was the FOE occurrence, i.e., if participants reported to have a FOE or not. According to the findings on error detection, identifying an error is followed by a behavioral slowdown (Logan & Crump, 2010; Rabbitt, 1966). We thus anticipated finding a general effect of error detection on the reaction time of the answer to the FOE question: FOE reports should be significantly slower than no FOE reports. We found that FOE reports were significantly faster after a triplet that was answered correctly compared to those that followed a triplet that was answered incorrectly (see Fig. 7A). Both bisectable ($t(27) = -6.2$, $p < 0.001$) and non-bisectable triplets ($t(27) = -6.2$, $p < 0.01$) showed this effect. Moreover, for the bisectable NBT, the occurrence of the FOE (if participants reported to have a FOE) had an effect on the reaction time of the FOE report – i.e., participants took longer to report they felt they had committed an error, mean ± SD = 1007 ± 207 ms, than to report not having a FOE, mean ± SD = 846.6 ± 254 ms, ($F(1,27) = 26.26$, $p < 0.001$, partial $\eta^2 = 0.49$, see Fig. 7B). Similarly, for the non-bisectable triplets, there was a significant effect of the occurrence of the FOE ($F(1,27) = 59.1$, $p < 0.001$, partial $\eta^2 = 0.68$) on the reaction time of the FOE report. Participants took longer to report that they had a FOE, mean ± SD = 1071 ± 387 ms, than to report that they did not have it, mean ± SD = 805 ± 240 ms, see Fig. 7B.

3.4. Confidence and the degree of awareness of missed errors

We explored participants' degree of awareness of erroneous responses in cases where they did not report an error (i.e., missed errors) by assessing if there were significant differences between the confidence ratings reported after triplets that were answered correctly and triplets that were answered incorrectly (but which were not described as being wrong, i.e., when participants did not report a FOE) (paired sample *t-test*). Note that confidence ratings were collected only when participants reported that they did not have a FOE (see section 2.3 for details). Descriptive statistics and box plots (see Fig. 8A and B) show that the distribution of the confidence ratings differed depending on whether the NBTs were answered correctly or incorrectly. That is, despite participants reported not having FOEs, they reported having both low and high levels of confidence spread all over the confidence scale. Moreover, in spite of not having a FOE, the confidence reported by the participants was significantly higher for all the objectively correct, mean ± SD = 5.1 ± 0.6, than for incorrect, mean ± SD = 3.4 ± 1.1, trials ($t = 9.25$, $df = 26$, $p < 0.001$, two tailed). This was also the case when bisectable ($t = 7.88$, $df = 25$, $p < 0.001$, two tailed) and non-bisectable ($t = 8.09$, $df = 20$, $p < 0.001$, two tailed) NBTs were considered separately. In general, participants were significantly more confident about their answers for the non-bisectable triplets, mean ± SD = 4.9 ± 0.9 than for the bisectable ones, mean ± SD = 4.2 ± 1.1, ($t = 4.8$, $df = 27$, $p < 0.001$, two tailed). This is also in line with the results reported above: non-bisectable triplets were shown to be the easier ones (see Section 3.1).

The fact that the confidence ratings were highly accurate about the correctness of the NBT motivated us to further explore its possible determinants. We thus performed analyses of the confidence ratings to assess if fluency while solving the NBT also influenced the observed variations in the confidence reports. Repeated measures ANOVAs with the same factors used for the FOE reports were performed to determine if fluency had an effect on the confidence ratings.

For the bisectable triplets, the factor *range* showed a significant effect on the confidence reports; participants were significantly more confident about their answer if the NBTs had a small range, i.e., fluent triplets, mean ± SD = 5.1 ± 0.8, compared to the ones that had wide range, i.e., disfluent triplets, mean ± SD = 4.2 ± 1.1, ($F(1,27) = 73.515$, $p < 0.001$, partial $\eta^2 = 0.731$). For the non-bisectable triplets the *bisection possibility* and the *distance to mean* had an effect on the confidence ratings. *Bisection possibility*; $F(1,27) = 7.305$, $p = 0.012$, partial $\eta^2 = 0.213$, *distance to mean*; $F(1,27) = 62{,}309$, $p < 0.001$, partial $\eta^2 = 0.698$. Confidence ratings were significantly higher for impossibility of bisection, i.e., fluent triplets, mean ± SD = 5.1 ± 0.7, than for triplets with possibility of bisection, i.e., disfluent ones, mean ± SD = 4.9 ± 0.7. Additionally,

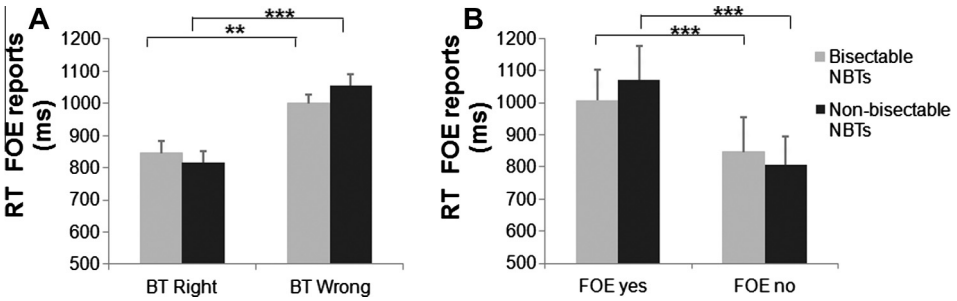

Fig. 7. Reaction time of FOE reports. **(A)** The reaction time of FOE reports reveals that post-decision evaluation processes occurred after incorrectly answered triplets compared to correctly answered ones. **(B)** Participants were slower in reporting a FOE versus reporting that they did not have a FOE. Error bars indicate the upper half of the 95% confidence interval, $^{**}p < 0.01$; $^{***}p < 0.001$.

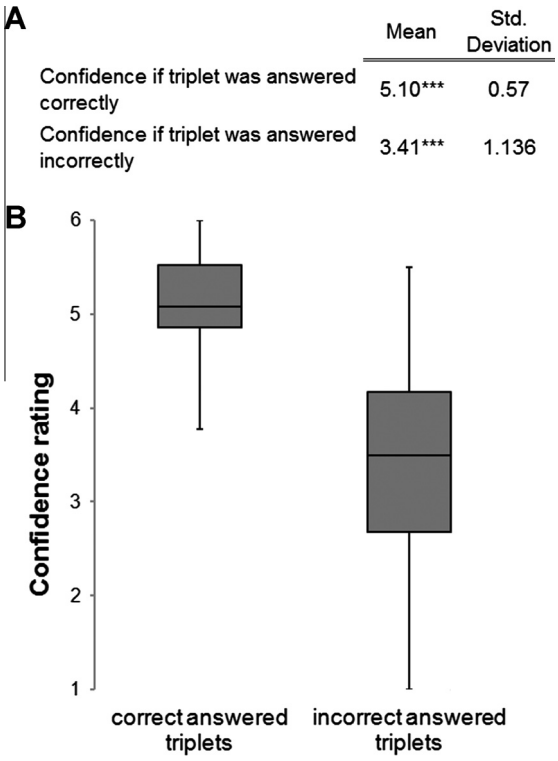

Fig. 8. Descriptive statistics of the confidence ratings. **(A)** Mean and standard deviation of the confidence ratings, $^{***}p < 0.001$. **(B)** Box plots of the confidence ratings from the correct and incorrect answered triplets.

the confidence ratings were higher for fluent triplets with larger *distance to mean*, mean ± SD = 5.4 ± 0.5, as compared with disfluent ones with smaller *distance to mean*, mean ± SD = 4.6 ± 0.9.

4. Discussion

We used a mathematical reasoning task paired with a speeded response paradigm to evoke mental reasoning errors, and then asked the subjects to report their FOEs and their confidence in the answers. This allowed us to integrate the literature on metacognition and error detection as well as their models and methods. All in all, we empirically showed (i) that the metacognitive FOE is a reliable signal of error, (ii) this FOE is determined by fluency (but the contribution of fluency as a determinant is limited to the general task difficulty), (iii) that post-decision evaluation mainly determines the occurrence of FOEs regardless of the difficulty of the task, and (iv) that participants' confidence in the correctness of their answers to the NBT was accurate even when they did not report a FOE.

4.1. Motor error detection versus mental error detection

We considered that the speeded decision normally imposed in the error detection paradigms was the method suited to study the spontaneity and immediateness of the FOEs. However, given the very specific types of well-defined paradigms routinely used in the error detection tradition, in which stimulus–response rules are simple (one-to-one) mappings known from the start and difficult to obey only due to time pressure and conflicting information (e.g., Flanker paradigm, Simon task), this approach seems unavoidably limited to the study of motor or bodily error detection. Errors in this sort of tasks are motor errors (e.g., inability to suppress inappropriate motor responses) rather than properly mental or reasoning-based errors. Hence, the related awareness studied so far in the error detection tradition has concerned the *awareness of motor errors* rather than *metacognitive awareness of mental actions and processes* (Proust, 2013).

In our study, we overcame this limitation of error detection studies by adopting a mathematical reasoning task in which the appropriate motor action is contingent upon a previous mental calculation, so that there is no simple mapping between the correct response and the motor action. In other words, our paradigm allowed us to actually test participants' monitoring of mental errors in mathematical reasoning.

Furthermore, literature on error detection has traditionally focused on the processing and consequences of errors (e.g., Dutilh et al., 2012), the mapping of errors with event-related potentials (e.g., Chang, Davies, & Gavin, 2009; Scheffers & Coles, 2000; Yeung & Summerfield, 2012, see also Wessel, 2012; for a review) and the neural correlates of error detection using fMRI (e.g., Ullsperger & Cramon, 2004). This focus has thus left the phenomenological dimension of the error detection, i.e., the FOE, out of the analysis. To bridge this gap, we resorted to the metacognitive tradition's method of asking participants to report their feelings and metacognitive states of awareness. We complemented this strategy by including participants' confidence ratings about the correction of their answers to the NBT. In this way, the present study fruitfully integrated the methods of both the error detection and the metacognitive traditions to address the way in which individuals detect and become aware of their mental reasoning errors.

4.2. The accuracy of the FOE

FOEs were successfully evoked in all 28 participants while they solved a mental calculation task. On average, each participant reported having a FOE in 21% of the NBTs that they solved, i.e., on average about 30 FOEs per participant. Notably, the FOEs

occurred throughout the experiment demonstrating that FOEs were triggered neither by the uncertainty of the initial trials nor by the tiredness of the final trials. We take this to indicate that FOEs were recurrent phenomena even when participants were able to perform the task well.

In line with other studies on metacognitive feelings (e.g., Gangemi & Bourgeois-Gironde, 2014; Paynter et al., 2009; Reder & Ritter, 1992) and performance monitoring (e.g., Fleming, Huijgen, & Dolan, 2012; Scheffers & Coles, 2000), our results provide evidence that the FOEs are very accurate. The resolution of the FOE (as measured by γ) had a mean value of 0.74 indicating a robust association between the objectively incorrect answers and the number of reported FOEs, as well as a strong relationship between the objectively correct answers and the lack of FOE reports. However, recent studies haven't found a correlation between metacognitive feelings and accuracy (Prowse Turner & Thompson, 2009; Thompson, 2014; Thompson et al., 2011, 2013).

The incongruence of our findings with those of Thompson et al. (2011, 2013) may be explained by a framing effect, i.e., the susceptibility of participants to the question or instruction wordings (Finn, 2008; Koriat, Bjork, Sheffer, & Bar, 2004). Particularly, the difference between the findings of Thompson and colleagues and ours may result from how the metacognitive question is asked: whereas they asked about the feeling of *rightness* and *certainty* (FOR-framed question), we asked for the feeling of *error* (FOE-framed question). As a result, a FOR-framed question may lead to lower accuracy and overconfidence as compared to a FOE-framed question. This effect is similar to the framing effect found in metamemory studies which showed that remember-framed questions lead to inaccuracy and overconfidence in the judgements of learning (JOLs) as compared with the forget-framed JOLs (Finn, 2008; Koriat et al., 2004). Since Thompson et al.'s (2011, 2013) studies framed the metacognitive question in terms of rightness and certainty, this might have cued the participants to overlook their reasoning errors. Thus, a systematic study comparing positively-framed and negatively-framed metacognitive assessments in reasoning tasks appears promising.

It is worth noting that the instructions and time restriction imposed in our paradigm ensured that the FOE report was the first response that quickly and intuitively came to participants' awareness. Given that participants had limited time to answer the NBT – limiting their ability to re-calculate or actively reflect on the performed mental calculation – and were given only 2 seconds to spontaneously report whether they felt they had committed an error, our results suggest that this type of feeling-based metacognition (Koriat, 2007) provided participants with accurate assessments of their ongoing cognitive processes without the necessity of effortful and analytical thinking.

4.3. The limits of fluency

It is a well-known fact that fluency plays a central role in the monitoring of memory retrieval as well as in the monitoring of reasoning (Kelley & Lindsay, 1993; Koriat, 1993; Oppenheimer, 2008). In our study, fluency – or the easiness with which triplets were solved – was operationalized considering 4 mathematical factors that have been shown to influence the reaction time of the NBT in previous studies (Nuerk et al., 2002) (see Section 5. Supplementary material). That is, fluency was quantified using the reaction time of the answers to the triplets in dependence of the mathematical factors that have been shown to determine reaction time and error probability in the NBT.

In the present experiment, *range* was one of the factors that determined the fluency of the bisectable triplets. However, disfluent triplets with larger range did not significantly evoke more FOEs than fluent ones with shorter range. In the case of non-bisectable triplets, the factors *distance to mean* and *bisection possibility* significantly determined the fluency with which triplets were solved (for correct and incorrect triplets analyzed together). In these cases, the number of FOEs significantly increased for disfluent triplets (i.e., those with small distance to mean and possibility of bisection). To disentangle fluency effects on the FOEs from error rates across conditions (i.e., to assess if FOEs were determined by differences in fluency and not by increased error rates in disfluent triplets), we also performed this analysis separately for trials in which the NBT was solved correctly and incorrectly. This analysis revealed that the factor *distance to mean* had an effect on the number of reported FOEs for objectively correct and incorrect triplets. The effect of the factor *bisection possibility*, however, showed a trend toward statistical significance for the correct answered triplets. Considering that the number of trials that were non-bisectable, correctly answered and with a FOE was only 3.2 on average, it is possible that the effect size of the *bisection possibility* (partial $\eta^2 = 0.16$, when all triplets were analyzed together) was not strong enough to reach statistical significance with this small number of trials.

In sum, our results showed that fluency determined the occurrence of FOEs in the non-bisectable triplets but not in the bisectable triplets. A possible interpretation of this finding is that fluency, as a determinant of the FOE, is limited to the task's general difficulty. As our results demonstrate, the bisectable triplets were in general more difficult to solve than non-bisectable ones: significantly fewer bisectable triplets were answered correctly compared to non-bisectable ones (see Section 3.1). Furthermore, participants were significantly more confident about their answers for the non-bisectable compared to the bisectable triplets (see Section 3.4). We thus hypothesize that the fluency with which participants solve a task may determine the FOEs up to the point where a ceiling effect is reached and the effort needed to solve the task disrupts the link between fluency and the FOE. Previous studies have shown that effects of fluency can be disrupted when the cause of low fluency is known (Oppenheimer, 2006) or the system processing fluency effects is presented with competitive information (Topolinski & Strack, 2010). In the case of the current study, the difficulty of solving the mathematical task within time restrictions might have caused a disruption of the effect of fluency on the FOEs. Thus, this disruption might have caused a decline in the accuracy of the FOEs. This hypothesis is supported by the fact that the number of FOEs was significantly larger and the gamma resolution significantly reduced in the bisectable triplets as compared to the non-bisectable triplets (see Section 3.2).

4.4. Post-decision evaluation

The second factor assumed to determine the FOEs, besides fluency, was post-decision evaluation processes. We found a post-error slowing in the subsequent task. Particularly, participants took significantly longer to respond to the FOE question after they have made an error on the NBT. This finding is in line with previous demonstrations that trial-to-trial slowdown can extend from a first task to a secondary task (Cho et al., 2009; Forster & Cho, 2014; Notebaert & Verguts, 2011). Secondly, according to our hypothesis, we found that participants were significantly slower in reporting that they had a FOE than in reporting that they did not have a FOE; and this was the case for both types of triplets (bisectable and non-bisectable).

We take these results to support the previous findings that a behavioral slowdown is linked to error detection (Jentzsch & Dudschig, 2009; Logan & Crump, 2010; Rabbitt, 1966). However, the phenomenon of post-error slowing has also been previously described as the product of other cognitive processes different than post-decision evaluation processes. For example, the cognitive

control account suggests that post-error slowing indexes cognitive control allocated to enhance successive performance (Gehring & Fencsik, 2001). The orienting account puts forward that behavioral slowing is the result of an orienting reaction toward an unpredicted and/or infrequent event, in this particular case, an error (Notebaert et al., 2009). Notably, in our study it is particularly difficult to disentangle the evaluation and the orienting accounts, since both errors and FOEs were infrequent. Although a dissociation between these three accounts is out of the scope of the present study, it was particularly interesting for the present study's goals to test if the occurrence of a FOE was associated with behavioral slowdown. Moreover, recent literature suggests that these accounts are not mutually exclusive (Danielmeier & Ullsperger, 2011), i.e., the three models together could explain the behavioral slowing observed after an error.

In summary, our results suggest that both fluency and post-decision evaluation processes are determinants of the FOE, i.e., both are complementary when evaluating one's own performance and contribute to determining the FOEs. The high accuracy of the FOEs found in our study is explained by the parallel interaction between the fluency and post-decision evaluation processes. We hypothesize that once the effect of fluency has reached its limits, post-decision evaluation processes keep monitoring the performance, allowing participants to detect their errors even when tasks become very difficult.

4.5. Confidence and the degree of awareness of missed errors

Finally, we were interested in the relationship between FOE and confidence. Participants were asked to rate the confidence in their answers to the NBT only in cases where they did not report having a FOE. Interestingly, our results show that in some cases where participants didn't report a FOE, they could still have low confidence. Moreover, their confidence was accurate insofar as it was significantly lower for the objectively incorrect answers than for the objectively correct ones. The accuracy of their confidence ratings suggests that participants implicitly detected their errors in the NBT, despite their lack of FOE reports.

These results are in accordance with previous error detection studies proposing that performance monitoring does not work in a binary fashion – where errors are either detected or not – but in a continuous and graded manner (Boldt & Yeung, 2015; Murphy et al., 2012; Scheffers & Coles, 2000). Thus, in cases when participants are forced to make a binary judgment, they must impose an arbitrary threshold that excludes some slightly detected errors (Steinhauser & Yeung, 2010). This way, confidence could be used as a measurement of subthreshold FOEs, in line with the growing literature that points to the fact that subjects implicitly detect (they are slightly aware of) some of their "unreported" errors (De Neys, 2012, 2014; De Neys & Glumicic, 2008; De Neys et al., 2011). Thus, the low confidence reported after an error in the NBT (without FOE report) suggests that the monitoring system accumulates accessible information about the "wrongness" of a response until a threshold is reached and a FOE is evoked.

Furthermore, we performed post hoc analyses to explore if the fluency with which triplets were solved was also a determinant of the confidence report. The three factors that were used to operationalize fluency in this study (i.e., *range*, *distance to mean*, and *bisection possibility*) showed significant effects on the confidence reported by the participants, whereby disfluent trials were associated with lower confidence ratings compared to fluent ones.

Thus, the present study highlights the important role of feeling-based metacognition for cognitive monitoring as a tool for uncovering inaccuracies in mental calculations. However, it remains to be determined whether the reported findings could be expanded to other real-world situations and to other cognitive activities.

Acknowledgements

We thank Kateryna Samoilova, Daniel Morales, Korbinian Moeller, Kourken Michaelian, Juan Pablo Bermúdez Rey, and the three anonymous reviewers for their helpful comments on earlier versions of the manuscript. We would also like to thank Asher Koriat for his advice regarding our experimental design, and Eric Schwitzgebel and Santiago Amaya for conceptual suggestions. This study was funded by the Werner Reichardt Centre for Integrative Neuroscience (CIN) at the Eberhard Karls University of Tübingen. The CIN is an Excellence Cluster funded by the Deutsche Forschungsgemeinschaft (DFG) within the framework of the Excellence Initiative (EXC 307).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.09.005.

References

Arango-Muñoz, S. (2014). The nature of epistemic feelings. *Philosophical Psychology, 27*(February 2015), 193–211.
Arango-Muñoz, S., & Michaelian, K. (2014). Epistemic feelings, epistemic emotions: Review and introduction to the focus section. *Philosophical Inquiries, 2*(1), 97–122.
Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68.
Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*(5), 372–382.
Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience, 35*(8), 3478–3484.
Brown, A. S. (2012). *The tip of the tongue state*. Taylor & Francis.
Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 325–337.
Chang, W. P., Davies, P. L., & Gavin, W. J. (2009). Error monitoring in college students with attention-deficit/hyperactivity disorder. *Journal of Psychophysiology, 23*, 113–125.
Cho, R. Y., Orr, J. M., Cohen, J. D., & Carter, C. S. (2009). Generalized signaling for control: Evidence from postconflict and posterror performance adjustments. *Journal of Experimental Psychology: Human Perception and Performance, 35*(4), 1161–1177.
Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology, 2*(September), 233.
De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*, 28–38.
De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*(January 2015), 169–187.
De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE, 6*(1), e15954.
De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*, 1248–1299.
De Sousa, R. (2009). Epistemic feelings. *Mind and Matter, 7*(2), 139–161.
Dhar, M., Wiersema, J. R., & Pourtois, G. (2011). Cascade of neural events leading from error commission to subsequent awareness revealed using EEG source imaging. *PLoS ONE, 6*(5), 19–22.
Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics, 74*(2), 454–465.
Endrass, T., Franke, C., & Kathmann, N. (2005). Error awareness in a saccade countermanding task. *Journal of Psychophysiology, 19*(2003), 275–280.
Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36*(4), 813–821.
Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision-making. *Journal of Neuroscience, 32*(18), 6117–6225.
Forster, S. E., & Cho, R. Y. (2014). Context specificity of post-error and post-conflict cognitive control adjustments. *PLoS ONE, 9*(3).
Gangemi, A., & Bourgeois-Gironde, S. (2014). Feelings of error in reasoning — In search of a phenomenon. *Thinking & Reasoning* (January 2015), 37–41.
Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2014). Feelings of error in reasoning — in search of a phenomenon. *Thinking & Reasoning*, 1–14 (ahead-of-print).
Gehring, W. J., & Fencsik, D. E. (2001). Functions of the medial frontal cortex in the processing of conflict and errors. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 21*(23), 9430–9437.

Hughes, G., & Yeung, N. (2011). Dissociable correlates of response conflict and error awareness in error-related brain activity. *Neuropsychologia, 49*(3), 405–415.
Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Quarterly Journal of Experimental Psychology, 62*(March 2015), 209–218.
Jiang, Y., & Hong, J. (2014). It feels fluent, but not right: The interactive effect of expected and experienced processing fluency on evaluative judgment. *Journal of Experimental Social Psychology, 54*, 147–152.
Kelley, C., & Lindsay, D. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1–2.
Koren, D., Seidman, L. J., Goldsmith, M., & Harvey, P. D. (2006). Real-world cognitive—and metacognitive—dysfunction in schizophrenia: A new approach for measuring (and remediating) more "right stuff". *Schizophrenia Bulletin, 32*(2), 310–326.
Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639.
Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9*(2 Pt 1), 149–171.
Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*(4), 643–656.
Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478–492.
Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness*. New York, USA: Cambridge University Press.
Logan, G. D., & Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science, 329*(October), 571–575.
Moeller, K., Klein, E., Fischer, M. H., Nuerk, H.-C., & Willmes, K. (2011). Representation of multiplication facts – Evidence for partial verbal coding. *Behavioral and Brain Functions: BBF, 7*, 25.
Moeller, K., Wood, G., Doppelmayr, M., & Nuerk, H. C. (2010). Oscillatory EEG correlates of an implicit activation of multiplication facts in the number bisection task. *Brain Research, 1320*, 85–94.
Moulin, C. J. A., & Souchay, C. (2013). Epistemic feelings and memory. In T. Perfect & S. Lindsay (Eds.), *Handbook of applied memory* (pp. 520–539). SAGE.
Murphy, P. R., Robertson, I. H., Allen, D., Hester, R., & O'Connell, R. G. (2012). An electrophysiological signal that precisely tracks the emergence of error awareness. *Frontiers in Human Neuroscience, 6*(March), 65.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133.
Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology, 38*, 752–760.
Notebaert, W., Houtman, F., Van Opstal, F., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition, 111*(2), 275–279.
Notebaert, W., & Verguts, T. (2011). Conflict and error adaptation in the Simon task. *Acta Psychologica, 136*, 212–216.
Nuerk, H., Geppert, B., van Herten, M., & Willmes, K. (2002). On the impact of different number representations in the number bisection task. *Cortex, 1*, 691–715.
Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology, 20*(October 2005), 139–156.
Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241.
Paynter, C. A., Reder, L. M., & Kieffaber, P. D. (2009). Knowing we know before we know: ERP correlates of initial feeling-of-knowing. *Neuropsychologia, 47*(3), 796–803.
Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford University Press.
Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning, 15*(March 2015), 69–100.
Rabbitt, P. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology, 71*(2), 264–272.
Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435.
Scheffers, M. K., & Coles, M. G. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance, 26*(1), 141–151.
Schwartz, B. L. (2001). *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval.. Experimental Psychology*. Psychology Press (pp. 239–240). Psychology Press.
Schwartz, B. L., & Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition, 39*(5), 737–749.
Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & R. E. Smith (Eds.), *The mind in context* (pp. 105–125). New York: Guilford.
Steinhauser, M., & Yeung, N. (2010). Decision processes in human performance monitoring. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(46), 15643–15653.
Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.
Thompson, V. (2014). What intuitions are... and are not. *Psychology of Learning and Motivation, 60*, 35–75.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*(January 2015), 215–244.
Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140.
Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*(2), 237–251. http://dx.doi.org/10.1016/j.cognition.2012.09.012.
Topolinski, S., & Strack, F. (2010). False fame prevented: Avoiding fluency effects without judgmental correction. *Journal of Personality and Social Psychology, 98*(5), 721–733. http://dx.doi.org/10.1037/a0019260.
Ullsperger, M., & Yves von Cramon, D. (2004). Neuroimaging of Performance Monitoring: Error Detection and Beyond. *Cortex*, 0010-9452, *40*(4–5), 593–604. http://dx.doi.org/10.1016/S0010-9452(08)70155-2.
Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgements: I. Properties of a self-regulating accumulator. *Nonlinear Dynamics, Psychology, and Life Sciences, 2*, 169–194.
Vickers, D., & Lee, M. D. (2000). Dynamic models of simple judgments: II. Properties of a self-organizing PAGAN (parallel, adaptive, generalized accumulator network) model for multi-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences, 4*(1), 1–31.
Wessel, J. R. (2012). Error awareness and the error-related negativity: Evaluating the first decade of evidence. *Frontiers in Human Neuroscience, 6*(April), 88.
Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences, 367*(1594), 1310–1321.